

**AI-Assisted feedback
for student revisions of short responses.**

Aubrey Condor
11/07/23

Why AI-assisted Feedback (AIF)?

- Research shows benefits of writing out explanations, and revising written responses
- Encouraging meaningful revision is challenging!
 - Students fix mechanical errors / fail to change anything
- Providing timely/useful feedback -> positive learning outcomes
 - In-person & computerized guidance



My Previous Work

- Automatic Grading -> Explainability (?)
- Trained an agent to correct an OE response
 - Large Language Model (LLM) + Deep Reinforcement Learning (RL)
 - Adds “key phrases” to a student’s response
 - representing a concept the student failed to include
 - Can also delete portions of the response
 - but almost always chooses to add a key phrase

Q: How can I use the RL agent to help a student improve their own response?

Feedback Design

Goals:

- 1) Provide actionable feedback
- 2) Provide suggestive advice
- 3) Elicit results that allow us to examine students' perceptions
- 4) Customizable to an educator's preferences

I compare feedback from the RL agent to:

- 1) feedback from ChatGPT
- 2) static feedback (control group)

Overview of the experiment

OATutor (v1.5.1)

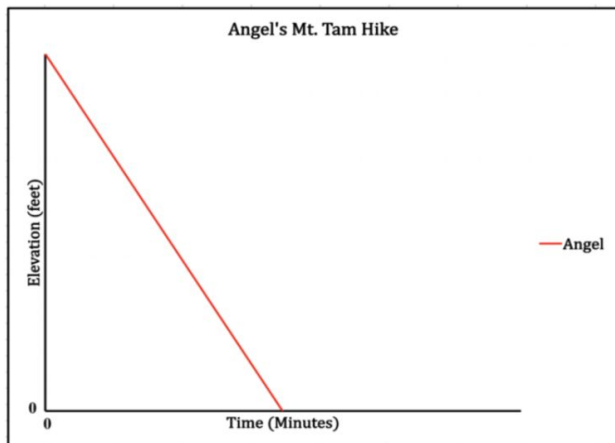
Elements of an Equation

Elements of an Equation

Angel is hiking down to the bottom of Mt. Tam, with an elevation of 2000 feet (ft).

She hikes at a constant rate of speed on her way down.

The graph below depicts Angel's hike (as a red line) from the top of Mt. Tam, 2000 ft., down to the bottom.



Students answer demographic & pretest items (from BEAR center)

Each symbol in the linear equation $y = mx + b$ represents a “real-life” component of Angel's 2000 ft hike to the bottom of Mt. Tam.

What does the symbol y represent?

- ☐ Total distance of Angel's hike.
- ☐ Angel's elevation after a given number of minutes.
- ☐ Number of minutes.
- ☐ Starting elevation of Angel's hike.
- ☐ Angel's change in elevation in one minute.

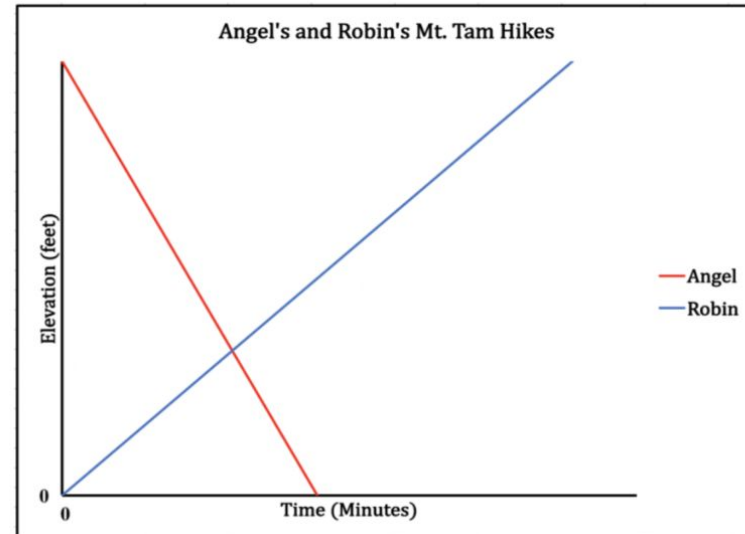
SUBMIT

Overview of the experiment

- 1) Students answer an OE item
 - mathematical problem solving from BEAR center
 - RL agent & Auto-grader trained on previously collected responses & ratings

Crossing Paths

Angel and Robin begin hiking at the same time, both at a constant rate of speed. Angel starts from the top of Mt. Tam, and Robin starts from the bottom. The red line shows Angel's downhill hike while the blue line shows Robin's uphill hike.



Overview of the experiment

2) RL & ChatGPT groups receive immediate feedback from an “AI bot”

- Students in the control group are given a static, one-sentence hint about the slope of a line

Who made it to their destination faster?

☐ Robin

☒ Angel

SUBMIT



What feature(s) of the graph led you to make your selection above?

We are training an Artificial Intelligence bot to change your response so that it is more correct! See how the bot revised your response below, then proceed to the next question.

it was because hiking downwards is faster than going up, a steeper line

It was because hiking downwards is faster than going up.

Overview of the experiment

3) Students critique the bot's revisions

4) Students engage in their own revision

- control group: asked if the hint they got was helpful, & prompted to revise

We need your help!

The bot is still in training so it doesn't write in perfect English, and sometimes messes up.

What did the bot add to, or remove from your response, and why does this make your response better or worse?

SUBMIT

Revise your own response.

Please keep in mind the AI bot's changes, and revise your own original response below.

SUBMIT

Research Questions

- 1) Is the RL AIF superior to AIF from ChatGPT and non-AI feedback in encouraging students to improve their original response?
- 2) Does the feedback intervention effect vary for students with different prior knowledge?
- 3) How do students perceive and act on the feedback they received, and do perceptions differ between intervention groups?
- 4) Can ChatGPT generate correct and useful revisions, and correspondingly what issues arise with using feedback from an unconstrained, generative AI?

RL Agent's Revisions - Quant Eval

- With one revision (either addition of a key phrase or removing a part of the response), an answer improves about 0.56 (std = 0.056) of a construct level
- On average, it takes the agent 2.3 (std = 1.37) revisions to achieve an expected score of 2.7 (where scores range from 0-3)

Example:

Original Student response: *"Angels line ended first on the x axis"*

Machine-revised response (RL): *"A steeper line, Angels line ended first on the x axis"*

RL Agent's & ChatGPT's Revisions - Qualitative Eval

7 SMEs evaluate the quality of machine revisions

See student responses and corresponding revisions, and respond on a Likert scale:

- 1) *“The machine-revised response is recognizable as the original student's response.”*
- 2) *“The machine-revised response includes a concept that was not present in the original student response.”*
- 3) *“The machine-revised response provides a clue about how the original response can be improved.”*

SME Results: Coded Likert Scale

Likert scale is ordinal!

0: strongly disagree, 7: strongly agree

<i>The machine-revised response ...</i>	<i>is recognizable as the original student's response.</i> (Statement 1)		<i>includes a concept that was not present in the original student's response.</i> (Statement 2)		<i>provides a clue about how the original response can be improved.</i> (Statement 3)	
	RL	GPT	RL	GPT	RL	GPT
mean	5.417	4.917	5.708	2.208	5.042	2.833
std	1.442	1.954	1.732	1.668	2.032	1.881

Student Data Collection

- **OATutor:** a Berkeley-created open-sourced system
- About 500 undergrads from University of Central Florida and State College of Florida
- Randomly allocated to control, RL and ChatGPT group
- Deleted records for participants who:
 - did not answer at least one demographic question
 - did not answer both the OE item and the revision item
 - <18 years old

group		age		gender		ethnicity		course	
ctrl	111	18-21	228	F	182	White	147	STA2023 (UCF)	191
rl	108	22-25	45	M	118	Hispanic	54	STA2023 (SCF)	42
gpt	110	25-30	21	Non binary	5	Asian	43	STA4102 (UCF)	39
		>30	10	Prefer not to say	4	Black/ AA	24	STA4173 (UCF)	23
						Multiple/ Other	32	STA1001 (UCF)	14
								MTH4420 (UCF)	4
n	329		304		309		300		313


Research Questions

- 1) Is the RL AIF superior to AIF from ChatGPT and non-AI feedback in encouraging students to improve their original response?**

Response Scores

OE: original student response

Revision: student-revised response

	All		Control		RL		GPT	
	OE	Revision	OE	Revision	OE	Revision	OE	Revision
n	329		111 		108		110	
mean	1.38	1.41	1.45	1.30	1.35	1.58	1.34	1.34
std	0.63	0.99	0.61	1.11	0.64	0.92	0.64	0.94

ANCOVA + Post Hoc pairwise

Source	SS	DF	F	p	np2
group	6.602	2	3.869	0.022	0.023
OE	45.609	1	53.454	0.000	0.141
Residual	277.3	325			

Comparison	Statistic	p-value	Lower CI	Upper CI
Control - RL	-0.282	0.000	-0.401	-0.163
Control - GPT	-0.046	0.629	-0.165	0.072
RL - Control	0.282	0.000	0.163	0.401
RL - GPT	0.236	0.000	0.116	0.355
GPT - Control	0.046	0.629	-0.072	0.165
GPT - RL	-0.236	0.000	-0.355	-0.116

Research Questions

2) Does the feedback intervention effect vary for students with different prior knowledge?

OLS fit with Prior Knowledge

	<u>coef</u>	Std err	t	P> t	[0.025	0.975]
Intercept	-0.660	0.419	-1.573	0.117	-1.485	0.165
Group-GPT	0.518	0.486	1.066	0.287	-0.438	1.474
Group-RL	1.420	0.520	2.733	0.007	0.398	2.443
pre_total	0.119	0.040	2.987	0.003	0.041	0.197
Group-GPT* pre_total	-0.030	0.048	-0.631	0.529	-0.124	0.064
Group-RL* pre_total	-0.105	0.051	-2.087	0.038	-0.205	-0.006
OE	0.509	0.082	6.238	0.000	0.349	0.670

Research Questions

3) How do students perceive and act on the feedback they received, and do perceptions differ between intervention groups?

Revising / Copy-pasting

How many students did NOT engage in revision?

- control: 19, RL: 10 group, GPT: 20

How many students did a copy-paste of machine-revision for their own?

- RL: 4, GPT: 8
- w/ cosine-similarity ≥ 0.98 : RL: 10, GPT: 10

Student Sentiment towards feedback

Manually coded all student critiques

- Positive: 48%
- Neutral: 35%
- Negative: 17%

For control, RL, and GPT groups respectively:

- Positive: 49%, 47%, 49%
- Neutral: 33%, 37%, 36%
- Negative: 19%, 16%, 15%

Positive: *“It clarified the slope to be the rate of change which made my response stronger”, “it added a better explanation”*

Negative: *“That made it incorrect because steeper does not mean more perpendicular.”, “The bot didn't use punctuation which made it harder to read.”*

Neutral: *“the bot took out my “mathematical” verbage”, “Changed less to fewer”*

More Specific Student Attitudes

Control group: 33% said they **would not use the information** provided

- Info already used it in their original response, or it wouldn't be helpful

ChatGPT and RL groups:

- 15% & 15% focused on **mechanistic changes** (grammar, punctuation, spelling)
 - *"It added a comma between line and the. Then it changed greater to faster. It is more or less the same answer."*
- 13% & 13% said the agent **didn't add/change anything**
- 40% & 40% said the bot's was **better** than their original:
 - *"It added context to the sentence. Made my answer better"*
- 7% RL and 12% GPT said revision was **worse**:
 - *"It removed angels name and it made the answer worse"*

Research Questions

4) Can ChatGPT generate correct and useful revisions, and correspondingly what issues arise with using feedback from an unconstrained, generative AI?

Analysis of ChatGPT Machine-Revisions

Manually reviewed ChatGPT revision & corresponding OE response

- **26% were incorrect**

- Student response: *“Positive slope”*
- ChatGPT revision: *“Robin made it to the destination faster because of the positive slope.”*

- 37% were correct and seemed helpful

- Student response: *“Angel's line finished first”*
- ChatGPT revision: *“Angel's line finished first because it has a steeper slope.”*

Analysis of ChatGPT Machine-Revisions

ChatGPT doesn't follow directions!

We tell ChatGPT to: *“respond only with your updated response”*

- 5% of cases, adds phrases like: *“the response should be:”, “The correct response to who made it to the destination faster and why is:”*

We tell ChatGPT to *“Please change no more than five words of my response”*

- 34% of cases, changes > 5 words:
 - Student response: *“The time in minutes”*
 - ChatGPT revision: *“Robin made it to the destination faster than Angel because the blue line has a positive slope, indicating an uphill hike, which means Robin was ascending while Angel was descending.”*

Analysis of ChatGPT Machine-Revisions

ChatGPT doesn't follow directions!

Sometimes (13%) it barely changes anything

- Student response: *"When robin reaches 0 elevation."*
- ChatGPT revision: *"When Robin reached 0 elevation,"*.

Or only swaps words for their synonyms (21%)

- Student response: *"The slope is steeper so it indicates less time."*
- ChatGPT response: *"The steeper slope suggests less time."*

6% of the time response with nonsense

- Student response: *"the x axis"*
- ChatGPT response: *"Robin made it to the destination faster and why"*

Limitations

- ChatGPT Prompt
- Manual review was just me
- Only used 1 OE item from a specific subject area
- Only undergraduates

